METHOD, PROGRAM AND APPARATUS FOR PREDICTING GENE EXPRESSION

BACKGROUND OF THE INVENTION

1) Field of the Invention

The present invention relates to a method, program and apparatus for predicting gene expression sites of a gene whose expression site is unknown on a genome sequence.

2) Description of the Related Art

While recent progress on the genetic engineering facilitates decoding of gene sequences, as the next approach, the functions of unknown genes on the genome sequences are analyzed. On analyzing genetic functions, prediction of the functions based on gene expression sites is effective. Therefore, a technique for predicting an expression site of unknown genes is required.

For predicting expression sites of an unknown gene, there are two methods. The first method is practically and experimentally exploring a tissue with expression site of the unknown gene. The second method is homology searching a large number of expressed sequence tag (hereinafter, "EST") sequences for unknown gene sequences, using a computer. This second method includes searching an EST database for a homology of gene sequences to be predicted and extracting expression information on the EST from the result of searching.

The first method has a problem in that functional analysis is not carried out efficiently because it depends on experiments. The second method has a problem in that fast functional analysis is prevented because it takes a longer time for management of a large number of EST sequences and the homology searching.

SUMMARY OF THE INVENTION

It is an object of the present invention to at least solve the problems in the conventional technology.

The method for predicting gene expression sites according to one aspect of the present invention includes calculating a distance between first and second genes on a genome sequence, wherein an expression site of the first gene is unknown, and the second gene is one of a plurality of genes whose expression sites are known; and determining the expression sites of the first gene based on the distance.

The computer program product according to another aspect of the present invention realizes the method according to the present invention on a computer.

The apparatus for predicting gene expression sites according to still another aspect of the present invention includes a calculation unit that calculates a distance between first and second genes on a genome sequence, wherein an expression site of the first gene is unknown, and the second gene is one of a plurality of genes whose expression sites are known; and a determination unit that determines the expression

sites of the first gene based on the distance.

The other objects, features and advantages of the present invention are specifically set forth in or will become apparent from the following detailed descriptions of the invention when read in conjunction with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is an example of arrangement of known genes and an unknown gene on a genome sequence;

Fig. 2 is a graph of the number of gene pairs having the same expression site versus distance between genes;

Fig. 3 is a hardware block diagram of an apparatus for predicting gene expression sites according to the embodiment of the present invention;

Fig. 4 is a functional block diagram of the apparatus for predicting gene expression sites;

Fig. 5 is a flowchart of procedures performed by the apparatus for predicting gene expression sites;

Fig. 6 is an example of sequence information on a sequence of an unknown gene;

Fig. 7 is an example of sequence information on a sequence of a known genome;

Fig. 8 is an example of comparison between the sequences of the known genome and the unknown gene;

Fig. 9 is an example of name and positional information of

known genes;

Fig. 10 is an example of computational result of distance between a known gene and an unknown gene;

Fig. 11 is an example of an expression profile;

Fig. 12 is an example of expression information of known genes;

Fig. 13 is another example of expression information of known genes;

Fig. 14 is a diagram showing a positional relation of genes on a genome sequence, from which the sensitivity and the specificity are derived;

Fig. 15 is a diagram (graph) showing computational results using the known data associated with the human chromosome 19;

Fig. 16 is a diagram (graph) showing computational results using the known data associated with the human chromosome 21;

Fig. 17 is a diagram showing the predicted results of the expression sites of the known gene (ABCC13);

Fig. 18 is a first diagram showing the predicted results of the expression sites of another known gene (Human gene GPR40); and

Fig. 19 is a second diagram showing the predicted results of the expression sites of another known gene (Human gene GPR40).

DETAILED DESCRIPTION

Exemplary embodiments of a method, a computer program product, and an apparatus relating to the present invention will be explained in detail below with reference to the accompanying drawings.

Fig. 1 is an example of arrangement of known genes and an unknown gene on a genome sequence. An unknown gene 101 is mapped on the genome sequence 100. Known genes 102 and 103, which are located around the unknown gene 101, are specified on the genome sequence 100. The known gene (hereinafter, "surrounding gene") 102 and the surrounding gene 103 are scored by distances (distances “a”, “b”) from the unknown gene 101. These distances are employed in prediction of expression sites of the unknown gene 101.

The surrounding gene 102 has expression sites in brain and ovary and has a distance “a” from the unknown gene 101. The surrounding gene 103 has an expression site in spleen and has a distance “b” from the unknown gene 101. As the distance “a” is less than the distance “b”, it can be predicted that the expression sites of the unknown gene 101 is more relevant to brain or ovary.

Fig. 2 is a graph of the number of gene pairs (human chromosome 21) having the same expression site versus distance (million base-pair) between genes. As can be seen from the graph, genes on the genome sequence 100 spaced by shorter distances from each other exhibit a trend to express on the same tissue. The method according to the embodiment is for predicting the expression sites of the unknown gene 101 based on this trend.

Fig. 3 is a hardware block diagram of an apparatus for predicting gene expression sites (hereinafter, "expression predicting apparatus") according to the embodiment of the present invention. The expression predicting apparatus shown in Fig. 3 includes a central

processing unit (hereinafter, "CPU") 301, a read only memory (hereinafter, "ROM") 302, a random access memory (hereinafter, "RAM") 303, a hard disk dive (hereinafter "HDD") 304, a hard disk (hereinafter, "HD") 305, a flexible disk drive (hereinafter, "FDD") 306, a flexible disk (hereinafter "FD") 307 as an exemplary detachable storage medium, a display 308, a communication interface (hereinafter, "I/F") 309, a keyboard 311, a mouse 312, a scanner 313, and a printer 314. These elements are connected to each other via a bus 300.

The CPU 301 controls the expression predicting apparatus. The ROM 302 stores programs such as a boot program. The RAM 303 is employed as a work area for the CPU 301. The HDD 304 controls read/write of data from/to the HD 305 under the control of the CPU 301. The HD 305 stores data written under the control of the HDD 304.

The FDD 306 controls read/write of data from/to the FD 307 under the control of the CPU 301. The FD 307 stores data written under the control of the FDD 306 and allows the expression predicting apparatus to read out the data stored in the FD 307. Other detachable storage media than the FD 307 may include a CD-ROM (CD-R, CD-RW), a magneto-optical disk (hereinafter "MO"), a digital versatile disk (hereinafter "DVD") and a memory card. The display 308 displays a cursor, icons, toolboxes, and data such as documents, images, and functional information. The display 308 is, for example, a cathode ray tube (CRT), a thin film transistor (TFT)-liquid crystal display or a plasma display.

The I/F 309 is connected to a network 31 such as a local area

network (LAN) or the Internet via a communication channel, and further connected via the network 310 to network nodes or server machines equipped with databases. The I/F 309 interfaces between the internal of the expression predicting apparatus and the network 310 to control input/output of data from/to the servers and the network nodes. The I/F 309 is, for example, a modem.

The keyboard 311 has a plurality of keys for entering characters, numerals and various instructions to input data. The keyboard 311 may be input pads of a touch panel type or a ten-key. The mouse 312 is employed to move the cursor, select a range to be processed, and move and resize a window. Pointing devices such as a track ball and a joystick may be employed, instead of the mouse 312.

The scanner 313 optically reads images such as graphics and pictures and sends them as image data to the expression predicting apparatus. The scanner 313 also has an optical character reader (hereinafter, "OCR") function to obtain data indicating contents of a document from printed one. The printer 314 prints image data and document data, and is, for example, a laser printer or an inkjet printer.

Fig. 4 is a functional block diagram of the expression predicting apparatus. An unknown gene sequence reader 401 reads sequence information of the unknown gene 101. The function of the unknown gene sequence reader 401 can be achieved using the I/F 309 or the keyboard 311, the mouse 312 and the scanner 313.

The genome sequence reader 402 reads information about the genome sequence 100 associated with the unknown gene 101. The

function of the gene sequence reader 402 can also be achieved using the I/F 309 or the keyboard 311, the mouse 312 and the scanner 313 similar to the unknown gene sequence reader 401.

The sequence comparator 403 compares the sequence information on the gene sequence 101 read by the unknown gene sequence reader 401 with the information on the genome sequence 100 read by the genome sequence reader 402. From the result of comparing, the sequence comparator 403 maps the sequence information on the unknown gene 101 onto the genome sequence 100. This mapping is detailed later.

The unknown gene position acquirer 404 acquires a comparison result 451 obtained from the sequence comparator 403, that is, a position on the genome sequence 100 of the unknown gene 100 mapped on the genome sequence 100. The acquired position is sent to the surrounding gene searcher 405 as positional information of unknown gene 452.

The surrounding gene searcher 405 searches for surrounding genes (genes 102, 103) having specified expression sites and located around the position of the unknown gene 101 on the genome sequence 100, based on name and positional information of surrounding gene 453 received from the genome sequence reader 402. As a result, the surrounding gene searcher 405 sends name information of surrounding gene 454 to the expression profile reader 406 and positional information of surrounding gene 456 to the surrounding gene expression site weighting processor 408.

From an expression profile database 553 later described, the expression profile reader 406 reads, under the control of the surrounding gene searcher 405, an expression profile corresponding to the surrounding gene name information.

From the expression profile read by the expression profile reader 406, the surrounding gene expression site acquirer 407 acquires the information on the expression sites specified by the surrounding gene and sends it as expression information of surrounding gene 455 to the surrounding gene expression site weighting processor 408.

The surrounding gene expression site weighting processor 408 computes distances on the genome sequence 100 between the surrounding genes 102, 103 having specified expression sites and a position of the unknown gene 101 to specify an expression site of the unknown gene 101 based on the computed distances. For example, the expression site(s) of one of more surrounding gene having shorter computed distances can be specified as the expression sites of the unknown gene 101.

The processor 408 computes distances between the surrounding genes having specified expression sites on the genome sequence information and a position of the unknown gene on the genome sequence. The processor 408 then sorts the surrounding genes in ascending order of computed distance. If a surrounding gene has the same expression information as that of the preceding surrounding gene in the sorted genes, the information is merged. In other words, the information about the same expression site is deleted to allow other

information than the information about the expression site to remain. An output unit 409 outputs only the remaining information other than the information about the same expression site.

As for the sequence comparator 403, the unknown gene position acquirer 404, the surrounding gene searcher 405, the surrounding gene expression site acquirer 407, and the surrounding gene expression site weighting processor 408, their functions can be achieved when the CPU 301 executes a program stored in the ROM 302, the RAM 303, the HD 305 or the FD 307 shown in Fig. 3.

The output unit 409 outputs the information about the expression sites processed by the surrounding gene expression site weighting processor 408 in a sorted-order of surrounding genes. The acquired information about the expression sites may be shown in a list, which corresponds to a predicted expression site list of unknown gene 457. The output unit 409 may allow the FD 307 and the I/F 309 shown in Fig. 3, to output the information to external, for example. It may also allow the printer 314 to print the information and the display 308 to display it.

Fig. 5 is a flowchart of procedures performed by the gene expression predicting apparatus. First, the gene expression predicting apparatus reads information of a sequence of an unknown gene to be predicted from a database 551 (Step S501). The unknown gene is hereinafter referred to as "target gene" and the sequence of the target gene as "sequence A". Next, information of genome sequence (hereinafter, "sequence B") is read from a database 552 and then is

mapped the sequence A onto the sequence B (Step S502). Further, the gene expression predicting apparatus calculates a distance between the target gene and a surrounding gene located around the target gene (Step S503). After that, an expression site of the surrounding gene is extracted from the expression profile database 553 (Step S504). Finally, the gene expression predicting apparatus weights the expression sites of the surrounding gene by the distance and outputs the weighted expression sites (Step S505).

These steps will be explained below in detail. Fig. 6 is an example of information of the sequence A read at Step S501. The information of the sequence A may be entered via the network 310 or a storage medium. In another method, it may be entered directly using the keyboard 311. In an alternative method, the scanner 313 having an OCR function may be employed to enter the information of the sequence A as image information, which is then converted into text data.

Fig. 7 is an example of information of the sequence B onto which the sequence A is mapped at Step S502. Fig. 8 is an example of comparison between the sequence B and the sequence A, and the comparison indicates the result of homology searching of the sequence A, that is, similar regions in the sequence B shown in Fig. 7. In other words, the comparison shown in Fig. 8 indicates the result of mapping at Step S502. The relation between the sequence A and the sequence B are required as follows: 1) the sequence A is mostly mapped onto the sequence B; 2) a gap 801 may be present in the sequence A which has

been mapped onto the sequence B; and 3) if the sequence A has the gap 801, each fragment of the sequence A divided by the gap 801 is mapped onto the sequence B in the order of appearance.

As can be seen from Fig. 8, a position (start position) of the sequence A, which is denoted as the reference number 802, on the sequence B is "12313789" (base-pair).

Fig. 9 is an example of name and positional information of surrounding genes around the sequence A, and the information is employed for calculating at Step S503. In Fig. 9, names of surrounding genes are described on the left column and positions (start positions) of the surrounding genes on the right column. The surrounding genes described are sorted in the order of appearance on the genome sequence 100. They may be sorted at the time of output in the step S505.

Fig. 10 is an example of computational results of distances between the sequence A and the surrounding genes, which are calculated by the following equation:

Distance (bp) between Sequence A of target gene and Surrounding gene = | (Position (bp) of Sequence A on Sequence B) – (Position (bp) of Surrounding gene on Sequence B) |

In Fig. 10, names of surrounding genes are described on the left column and distances of the surrounding genes from the target gene on the right column. The surrounding genes described are sorted in ascending order of the distance. For example, as can be seen from Fig. 9, the position of a surrounding gene "C21orf42" on the sequence B

is "12337804"(bp), and the position of the sequence A of the target gene on the sequence B is "12313789" (bp). Therefore, for this example, the distance from the target gene is "| 12313789 - 12337804 | = "24015" (bp). For another example, as can be seen from Fig. 9, since a surrounding gene "ADAMTS 1" has a position on the sequence B at "13788256" (bp), its distance from the target gene is "| 12313789 - 13788256 | = "1474467" (bp).

As a result, returning to Fig. 1, the distance between the surrounding gene 102 or 103 whose expression sites are known and the unknown gene 101 is calculated based on the start positions of both genes on the genome sequence 100. However, the start point of the distance may be a position other than the start position. For example, the distance between the both genes on the genome sequence 100 may be calculated as follows: 1) a distance between the end position of the unknown gene 101 and the end position of the surrounding gene; 2) a distance between the start position of the unknown gene 101 and the end position of the surrounding gene, without depending on the order of the surrounding gene and the unknown gene 101 on the genome sequence 100; 3) a distance between the end position of the unknown gene 101 and the start position of the surrounding gene, without depending on the order of the surrounding gene and the unknown gene 101 on the genome sequence 100; 4) a distance between any middle position (e.g. center) between the start and end positions of the unknown gene 101 and any middle position (e.g. center) between the start and end positions of the surrounding gene; 5) a distance between

any middle position (e.g. center) between the start and end positions of the unknown gene 101 and the start position of the surrounding gene; 6) a distance between any middle position (e.g. center) between the start and end positions of the unknown gene 101 and the end position of the surrounding gene; 7) a distance between the start position of the unknown gene 101 and any middle position (e.g. center) between the start and end positions of the surrounding gene; and 8) a distance between the end position of the unknown gene 101 and any middle position (e.g. center) between the start and end positions of the surrounding gene.

Fig. 11 is an example of an expression profile employed for extracting at Step S504, and the expression profile indicates contents of a surrounding gene "ADAMTS 1". In the item field "GENE" denoted by reference number 1101, an abbreviated name of the gene is recorded. In the item field "EXPRESS" denoted by reference number 1102, information about the expression sites are recorded. Other items are omitted to explain because they are not directly relevant to the embodiment.

Fig. 12 is an example of expression information of the surrounding genes shown in Fig. 6, and the expression information indicates the result of extracting at Step S504. In Fig. 12, names of surrounding genes are described on the left column and contents of the expression sites on the right column. The surrounding genes are sorted in ascending order of the distance and described as similar to Fig. 10. They may be sorted at the time of output in Step S505.

In the list shown in Fig. 12, expression sites in lower order are deleted from the list if they are the same as expression sites in higher order. For example, as shown in Fig. 12, the surrounding gene "MRPL39" located in lower order than the surrounding gene "C21orf42" in the most vicinity of the unknown gene has an expression tissue "testis". Though, it is found that the surrounding gene "C21orf42" has already expressed at the expression tissue "testis". Therefore, the expression site is deleted from the item of the lower surrounding gene "MRPL 39". Such operations are repeatedly performed to the lowermost surrounding gene. A deleted result is shown in Fig. 13.

After Steps S501 to S505 are completed, the list is obtained as shown in Fig. 13 with respect to the sequence A of the unknown gene. From this list, an expression site of the sequence A of the unknown gene can be predicted easily.

According to the embodiment as described, from the sequence of the unknown gene, its expression sites can be predicted easily. This is effective to reduce experimental costs and computer resources. Therefore, the use of the present invention makes it possible to promptly perform rapidly functional analysis of unknown genes, and can greatly contribute the development of the gene engineering.

In the present embodiment, as shown in Fig. 2, the expression sites of the unknown gene 101 is predicted based on the trend in that a given gene expresses on the same tissue as its surrounding genes (genes 102, 103). Therefore, the longer the distance between the unknown gene 101 and the surrounding gene (genes 102, 103), the

lower the precision of prediction drops. To maintain the precision of prediction, the distance cut off by a threshold is employed for predicting gene expression sites. The threshold is determined based on the sensitivity and specificity.

The sensitivity indicates a ratio of expression sites predicted (i.e. extracted) to expression sites previously determined that it is where the unknown gene expresses by another method. The higher the sensitivity, the less the pseudo-negative result arises and the higher the precision of prediction improves. The specificity indicates a ratio of expression sites not predicted to expression sites previously determined that it is where the unknown gene never expresses by another method. The higher the specificity, the less the pseudo-positive result arises and the higher the precision of prediction improves.

Therefore, based on the ratio (sensitivity) of expression sites extracted at Step S504 in Fig. 5, among expression sites previously determined by another method, on which the unknown gene 101 expresses, and on the ratio (specificity) of expression sites not extracted at Step S504, among expression sites previously determined by another method, on which the unknown gene never expresses, the threshold of the distance calculated at Step S503 is determined. Then, only the surrounding genes located within the determined threshold from target gene are to be sorted (Step S505). The Step S505 has the same process as described above.

Table 1 is a relationship between the sensitivity and the

specificity.

Table 1

	Number of Expression sites of Unknown gene Determined by Another Method	Number of Non-Expression sites of Unknown gene Determined by Another Method
Number of Expression sites Predicted	A	a
Number of Expression sites Not Predicted	B	b

The sensitivity and the specificity can be derived from the following calculations:

Sensitivity = A/(A + B)

Specificity = b/(a + b)

Fig. 14 is a diagram showing a positional relation of genes on a genome sequence, from which the sensitivity and the specificity are derived. Through the use of the information on the known gene, as shown in Fig. 14, genes 1401, 1402, and 1403 can be mapped onto a genome sequence 1400. To apply the predicting method to these genes, the gene 1401 is assumed as an unknown gene, and the genes 1402 and 1403 as surrounding genes. The same method is performed for the genes 1402 and 1403. Thresholds of the distance between the unknown gene and the surrounding gene are determined from, for example, an initial value of 100 kilo base-pair to 3000 kilo base-pair at an interval of 100 kilo base-pair. Averages of sensitivity and specificity calculated for each unknown gene within thresholds are calculated. Then, the calculated averages are plotted on a coordinate plane, which has the axis of abscissas indicating threshold between unknown and

surrounding genes and the axis of ordinates indicating sensitivity/specificity.

Fig. 15 is a diagram (graph) showing computational results using the known data associated with the human chromosome 19. Fig. 16 is a diagram (graph) showing computational results using the known data associated with the human chromosome 21. The axis of abscissas indicates threshold and the axis of abscissas indicates sensitivity/specificity both in Figs. 15 and 16. In Fig. 15, the longer the distance between unknown and surrounding genes, the lower the specificity drops and the higher the sensitivity elevates to the contrary. When a cross-point between the specificity and the sensitivity is employed as a threshold, pseudo-negative and pseudo-positive results can be reduced. This is possibly effective to maintain the precision of prediction. A threshold in Fig. 15 is found on 100 kilo base-pair, for example. Similarly, a threshold in Fig. 16 is found on 200 kilo base-pair, for example.

Thus, the threshold depends on chromosome. Accordingly, it is desirable to find respective thresholds for all chromosomes before prediction through the above calculations. The following results show the predicted expression sites using the threshold derived from the sensitivity and specificity.

EXAMPLE 1

Human gene ABCC13 is employed to predict expression sites. Fig. 17 is a diagram showing the predicted results of the expression

sites of the known gene (ABCC13). Human gene ABCC13, present on the human chromosome 19, is known in which its expression site is "Liver and Spleen" by the public database (UniGene).

On the other hand, as shown in Fig. 17, the surrounding gene "STCH" has an expression site "Liver and Spleen" denoted by reference number 1701, and this means that the expression site is correctly predicted. The paper (Biochem Biophys Res Commun. 2002 Dec 6; 299 (3): 410-7) confirms the expression sites of the human gene ABCC13 on the expression sites underlined in Fig. 17.

EXAMPLE 2

Human gene GPR40 is employed to predict expression sites. Figs. 18 and 19 are diagrams showing the predicted results of the expression sites of another known gene (Human gene GPR40). Human gene GPR40, present on the human chromosome 21, is known in which its expression site is "Islets of Langerhans" in human pancreas by the public database (UniGene).

On the other hand, as shown in Fig. 19, the surrounding gene "USF2" has an expression site "Islets of Langerhans" denoted by reference number 1901, and this means that the expression site is correctly predicted. The paper (J Biol Chem. 2003 Mar 28; 278 (13): 11303-11) confirms the expression sites of the human gene GPR40 on the expression sites underlined in Figs. 18 and 19.

The method for predicting gene expression sites according to the embodiment may be achieved by a previously prepared

computer-readable program, which is executed in computer such as a personal computer and a workstation. The program is stored in a computer-readable storage medium such as a HD, a FD, a CD-ROM, an MO and a DVD, and is read out of the storage medium by a computer to execute it. The program may include a transmission medium that can be delivered via a network such as the Internet.

As described above, genes on the genome sequence spaced by shorter distances from each other exhibit a trend to express on the same tissue. According to the invention utilizing this trend, it is possible to estimate an expression site of the unknown gene from information on the expression sites of the surrounding gene. This is effective to achieve a method, program and apparatus for predicting gene expression sites, which is possible to execute quick and efficient expression sites prediction and functional analysis for an unknown gene.

Although the invention has been described with respect to a specific embodiment for a complete and clear disclosure, the appended claims are not to be thus limited but are to be construed as embodying all modifications and alternative constructions that may occur to one skilled in the art which fairly fall within the basic teaching herein set forth.

SEQUENCE LISTING

<110> Fujitsu Limited

<120> A method, program and device for supporting prediction of gene ex